Clustered Stochastic Optimization for Object Recognition and Pose Estimation *

Juergen Gall, Bodo Rosenhahn, and Hans-Peter Seidel

Max-Planck-Institute for Computer Science,
Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany
{jgall, rosenhahn, hpseidel}@mpi-inf.mpg.de

Abstract. We present an approach for estimating the 3D position and in case of articulated objects also the joint configuration from segmented 2D images. The pose estimation without initial information is a challenging optimization problem in a high dimensional space and is essential for texture acquisition and initialization of model-based tracking algorithms. Our method is able to recognize the correct object in the case of multiple objects and estimates its pose with a high accuracy. The key component is a particle-based global optimization method that converges to the global minimum similar to simulated annealing. After detecting potential bounded subsets of the search space, the particles are divided into clusters and migrate to the most attractive cluster as the time increases. The performance of our approach is verified by means of real scenes and a quantative error analysis for image distortions. Our experiments include rigid bodies and full human bodies.

1 Introduction

Finding the 3D position and rotation of a rigid object in a set of images from calibrated cameras without any initial information is a difficult optimization problem in a 6-dimensional space. The task becomes even more challenging for articulated objects where the dimensionality of the search space is much higher, e.g., a coarse model of a human skeleton has already 24 degrees of freedom (DoF) yielding a 30-dimensional space. Although the initial pose is essential for many state-of-the-art model-based tracking algorithm, e.g. [1–3], relatively little attention was paid to the initialization of rigid and articulated models. A manual initialization is usually required, which is time demanding and assumes some expertise on the model and on the world coordinate system.

Depending on the image features, there are several techniques for pose estimation in the literature. Edge-based approaches, e.g. [4–6], align curves or lines of the model to detected edges. They work best for homogeneous objects, however, textured objects and cluttered background typically involve many edges that are not related to the model. Texture-based approaches [7, 8] use correspondences between the textured model and an image for pose estimation. Separate from the

* Our research is funded by the MPC for Visual Computing and Communication.

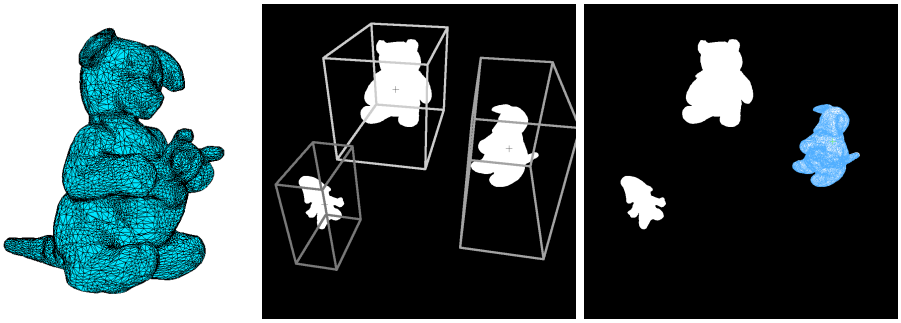

Fig. 1. From left to right: *a)* 3D model of object. *b)* Potential bounded subsets of the search space. *c)* Projection of the mesh. The pose is correctly estimated.

fact that they require textured surfaces for self-initialization, the texture needs to be registered to the model beforehand, i.e., a manual initialization is done for the texture acquisition during preprocessing.

Our approach for solving the initialization problem estimates the pose of rigid and articulated objects by minimizing an energy function based only on the silhouette information. Although we are not restricted to silhouettes, the object region has the advantage that it is an appearance independent feature that can be easily extracted from a single frame, e.g. by background subtraction. Since an initial guess is not available, local optimization algorithm like iterative closest point (ICP) [9, 10] are not suitable for this task. For finding the exact pose, we use a novel particle-based global optimization, called *interacting simulated annealing* [11], that converges to the global optimum similar to simulated annealing [12]. In order to deal with multiple objects, we extend the work in [11] by clustering the particles with respect to previously detected bounded subsets of the search space.

After a brief introduction to interacting simulated annealing in Section 2, we give details of our method in Section 3. In Section 4, some extensions for human bodies are explained. The experimental results are discussed in Section 5 followed by a brief conclusion.

2 Interacting Simulated Annealing

Interacting particle systems are well-known as particle filter [13] and approximate a distribution of interest η_t by $\eta_t^n := \sum_{i=1}^n \pi^{(i)} \delta_{X^{(i)}}$, where δ is the Dirac measure and $X^{(i)}$ are n random variables, termed particles, weighted by $\pi^{(i)}$. In the case of *interacting simulated annealing (ISA)*, the distribution is proportional to a Boltzmann-Gibbs measure

$$g_t(dx) = \exp\left(-\beta_t \, V(x)\right) \, \lambda(dx), \tag{1}$$

where $V \geq 0$ is the energy function to minimize, β_t is an annealing parameter that increases with t, and λ is the Lebesgue measure.

Algorithm 1 Interacting Simulated Annealing Algorithm

1. Initialization
   - Sample $x_0^{(i)}$ from η_0 for all i
2. Selection
   - Set $\pi^{(i)} \leftarrow \exp(-\beta_t V(x_t^{(i)}))$ for all i
   - For i from 1 to n:
     Sample κ from $U[0,1]$
     If $\kappa \leq \epsilon_t \pi^{(i)}$ then
     ⋆ Set $\check{x}_t^{(i)} \leftarrow x_t^{(i)}$
     Else
     ⋆ Set $\check{x}_t^{(i)} \leftarrow x_t^{(j)}$ with probability $\frac{\pi^{(j)}}{\sum_{k=1}^n \pi^{(k)}}$
3. Mutation
   - Sample $x_{t+1}^{(i)}$ from $K_t(\check{x}_t^{(i)}, \cdot)$ for all i and go to step 2

In contrast to particle filter that estimate the posterior distribution for a sequence of images, we apply ISA for estimating the global optimum in still images where no initial information is available. For this purpose, the steps *Selection* and *Mutation* of Algorithm 1 are iterated until the global minimum of V is well approximated. During the selection, the particles are weighted according to a given energy function V where greater weight is given to particles with a lower energy. The weights associated to the particles refer to the probability that a particle is selected for the next step. We used the parameter $\epsilon_t = 1/\sum_{k=1}^n \pi^{(k)}$ for selection since it has slightly better convergence properties than $\epsilon_t = 0$, see for instance [14, 11]. If a particle is not accepted with probability $\epsilon_t \pi^{(i)}$, a new particle is selected from all particles, e.g. by multinomial sampling. The selection process removes particles with a high energy while particles with a low energy are reproduced each time they are selected. An overview of various resampling schemes can be found in [15]. In the second step, the selected particles are distributed according to Markov kernels K_t specified by a modified dynamic variance scheme, which we propose in Section 3.4.

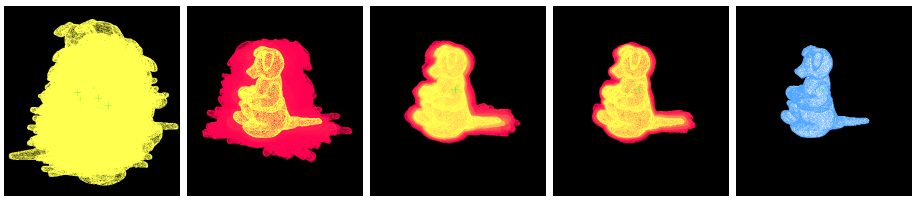

Fig. 2. Particles at $t = 0, 5, 10, 15$ and 19 for ISA. Particles with a higher weight are brighter, particles with a lower weight are darker. The particles converge to the pose with the lowest energy as t increases. **Most left:** Equally weighted particles after initialization. **Most right:** Estimate after 20 iterations.

While the annealing scheme prevents the particles from getting stuck in local minima, the dynamic variance scheme focuses the search around selected particles. When t increases only particles with low energy are selected and the search is concentrated on a small region, see also Figure 2. Indeed, it has been shown that ISA approximates a distribution η_t that becomes concentrated in the region of global minima of V as t tends to infinity provided that the annealing scheme β_t increases slow enough and the search space is bounded [16]. In [11], the authors evaluated several annealing schemes and parameter settings. In our experiments, a polynomial scheme, i.e.

$$\beta_t = (t+1)^b \quad \text{for some } b \in (0,1), \tag{2}$$

performed well with $b = 0.7$.

3 Clustered Optimization

3.1 Initial Subsets

Having a binary image for each camera view, where pixels that belong to the foreground are set to 1 else to 0, the pixels are first clustered with repsect to the 8-neighbor connectivity. In order to make the system more robust to noise, clusters covering only a very small area are discarded. In the next step, the 4 corners of the bounding box of each cluster are determined and the projection ray for each corner is calculated. The projection rays are represented as Plücker lines [17], i.e., the 3D line is determined by a normalized vector d and a moment m such that $x \times n = m$ for all x on the line. Provided that two projection rays from different views are not parallel, the midpoint p of the shortest line segment between the two rays l_1 and l_2 is unique and can be easily calculated. If the minimum distance between l_1 and l_2 is below a threshold, p is regarded as a corner of a convex polyhedron. After 8 corners of the polyhedron are detected for two clusters from two different views, the bounding cube is calculated as shown in Figure 1 b). In the case of more than two available camera views, each pair of images – starting with the views containing the most clusters – is checked until a polyhedron is found. The corners are similarly refined by calculating the midpoint of the shortest line segment between a ray from another view and a corner of a polyhedron. The resulting bounding cubes provide the initial bounded subsets of the search space. We remark that the algorithm is not very sensitive to the thresholds as long as the searched object is inside a bounding cube. This can be achieved by using very conservative thresholds.

3.2 Particles

Since we know the 3D model, the pose is determined by a vector in $\mathbb{R}^{6+m}$, i.e., each particle is a $6+m$-dimensional random vector where m is the number of joints. The rigid body motion M is represented by the axis-angle representation given by the 6D vector $(\theta\omega, t)$ with $\omega = (\omega_1, \omega_2, \omega_3)$ and $\|\omega\|_2 = 1$. The mappings

from $\theta\omega$ to a rotation matrix R and vice versa can be efficiently computed by the Rodriguez formula [18] and are denoted by $\exp(\theta\omega)$ and $\log(R)$, respectively.

Since ISA approximates a distribution by finite particles, we take the first moment of the distribution as estimate of the pose, i.e., the mean of a set of rotations $r^{(i)}$ weighted by $\pi^{(i)}$ is required.[1] This can be done by finding a geodesic on the Riemannian manifold determined by the set of 3D rotations. When the geodesic starting from the mean rotation in the manifold is mapped by the logarithm onto the tangent space at the mean, it is a straight line starting at the origin, see [19]. The tangent space is called exponential chart. Hence, the weighted mean rotation $\bar{r}$ satisfies

$$\sum_i \pi^{(i)} \left(\bar{r}^{-1} \star r^{(i)} \right) = 0, \tag{3}$$

where $r^{(j)} \star r^{(i)} := \log\left(\exp(r^{(j)}) \cdot \exp(r^{(i)})\right)$ and $r^{-1} := \log\left(\exp(r)^T\right)$. The weighted mean can thus be estimated by

$$\hat{r}_{t+1} = \hat{r}_t \star \left(\frac{\sum_i \pi^{(i)} \left(\hat{r}_t^{-1} \star r^{(i)} \right)}{\sum_i \pi^{(i)}} \right). \tag{4}$$

3.3 Initialization

Due to multiple objects as shown in Figure 1, each particle belongs to a certain cluster C given by the bounding cubes and denoted by $x^{(i,C)}$. At the beginning, a small number of particles is generated with different orientations located in the center of the cube for each cluster. The complete set of particles is initialized by randomly assigning each particle the values of one of the generated particles. Afterwards, each particle is independently diffused by a normal distribution with mean $x^{(i,C)}$ and a diagonal covariance matrix with fixed entries except for the translation vector t where the standard deviations are given by the edge lengths of the cube divided by 6 such that over 99.5% of the particles are inside the cube.

3.4 Mutation

The dynamic variance scheme for the mutation step is implemented by cluster dependent Gaussian kernels $K_t^{(C)}$ with covariance matrices $\Sigma_t^{(C)}$ proportional to the sampling covariance matrix of each cluster:[2]

$$\Sigma_t^{(C)} := \frac{d}{|C|-1} \sum_{\substack{i=1 \\ i \in C}}^{n} (x_t^{(i,C)} - \mu_t)_\rho \, (x_t^{(i,C)} - \mu_t)_\rho^T, \quad \mu_t := \frac{1}{|C|} \sum_{\substack{i=1 \\ i \in C}}^{n} x_t^{(i,C)}, \tag{5}$$

[1] The density could also be estimated by kernel smoothing from the particles in order to take the peak of the density function as estimate. However, kernel smoothing is more expansive than calculating the first moment of a density and it also needs to be performed in the space of 3D rotations.

[2] Samples from a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ can be drawn via a Cholesky decomposition $\Sigma = AA^T$: $x = \mu + Az$ where z is drawn from $\mathcal{N}(0, I)$.

where $|C|$ is the number of particles in cluster C and $((x)_\rho)_k = \max(x_k, \rho)$ for the k^{th} dimension. The value $\rho > 0$ ensures that the variance does not become zero for any dimension. In practice, we set $d = 0.4$ and compute only a sparse covariance matrix, see also Section 4.

3.5 Selection

Since each particle defines the pose of the model, the fitness of a particle $x \in \mathbb{R}^{6+m}$ can be evaluated by the difference between the original image and the template image that is the projected surface of the model. For this purpose, we apply a signed Euclidean distance transformation [20] on the silhouette image I_v and on the template $T_v(x)$ for each view v. The energy function is defined by $V(x) := \frac{\alpha}{r} \sum_{v=1}^{r} V_v(x)$ with

$$V_v(x) := \frac{1}{2|T_v^+(x)|} \sum_{p \in T_v^+(x)} |T_v(x,p) - I_v(p)| + \frac{1}{2|I_v^+|} \sum_{p \in I_v^+} |T_v(x,p) - I_v(p)|, \quad (6)$$

where I^+ denotes the set of strictly positive pixels of an image I. The normalization constant $\alpha = 0.1$ ensures that V is approximately in the range between 0 and 10, which is suitable for the selected annealing scheme.

The resampling step is cluster independent, i.e., the particles migrate to the most attractive cluster where the particles have more weight and give more offspring. At the end, there are no particles left where the silhouettes do not fit the model, see Figure 1 c).

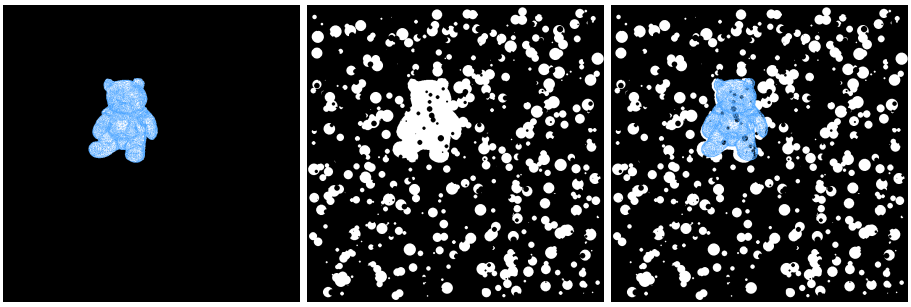

Fig. 3. From left to right: *a)* Estimated pose without noise. The error is less than $1mm$ (median). *b)* Silhouettes are randomly distorted by 500 white and 500 black circles. *c)* Median estimate with error less than $4cm$.

4 Human Bodies

While for rigid bodies the correlation between the parameters is neglected due to computational efficiency, correlation between connected joints in the human

skeleton are incorporated. That is, the correlation of the joints that belong to the same skeleton branch, e.g. the left leg, are calculated in the dynamic variance scheme (5) while correlations with joints to other branches are set to zero.

In order to focus the search on poses with higher probabilities, prior knowledge is incorporated into the energy function as soft constraint. The probability of a pose p_{pose} is estimated by a Parzen-Rosenblatt estimator with Gaussian kernels [21, 22] over a set of subsamples from different motions from the CMU motion database [23]. Since the dependency between the joints of the upper body and the joints of the lower body is low, the sample size can be reduced by splitting p_{pose} up into two independent probabilities p^u_{pose} and p^l_{pose}, respectively. Hence, the energy function is extended by

$$V(x) := \frac{\alpha}{r} \sum_{v=1}^{r} V_v(x) - \frac{\eta}{2} \ln \left(p^u_{pose}(x) p^l_{pose}(x) \right), \tag{7}$$

where $\eta = 2.0$ regulates the influence of the prior. Moreover, the mean and the variance of the joints in the training data is used to initialize the particles. To get rid of a biased error from the prior, the final pose is refined by ICP [9, 10] that is initialized by the estimate of ISA.

5 Results

For the error analysis, synthetic images with silhouettes of the bear were generated by projecting the model for 3 different views. The error was measured by the Euclidean distance between the estimated 3D position and the exact position. Each simulation was repeated 25 times and the average errors for different numbers of particles and iterations are plotted in Figure 5. The estimates for 200 particles and 30 iterations are very accurate with a median error less than $1mm$, see Figure 3 a). The influence of distorted silhouettes is simulated by randomly drawing first a fixed number of white circles and then black circles. Holes, dilatation and erosion are typically for background subtraction and change the outcome of the Euclidean distance transform. The diagrams in Figure 5 show that our method performs also well for distorted silhouettes. In the case of 500 white and 500 black circles, the error of the median estimate shown in Figure 3 is still less than $4cm$. The performance for a human body with 30 DoF was tested by generating synthetic images with silhouettes for 12 single poses from a sequence of the CMU database that was not used for the prior. The estimates are given in Figure 6. The average error of the joints for 400 particles and 40 iterations was 1.05°. Results for a real scene with background subtraction are shown in Figure 4. For images of size 1004×1004 pixels, the computation cost is given by number of views $\times$ number of iterations $\times$ number of particles $\times$ 0.0346 seconds.

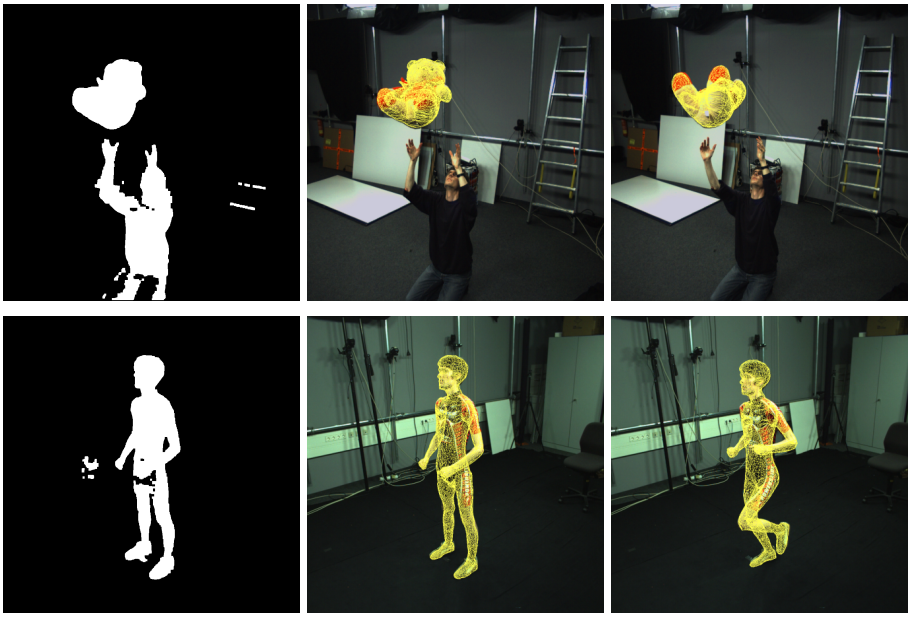

Fig. 4. Estimates for a real scene. 3 views were segmented for the bear and 4 views for the human (Only one is shown). **Most left:** Silhouettes from background subtraction.

6 Discussion

We proposed an accurate and robust approach, which relies on a global optimization method with clustered particles, for estimating the 3D pose of rigid and articulated objects with up to 30 DoF. It does not require any initial information about position or orientation of the object and solves the initial problem as it occurs for tracking and texture acquisition. Our experiments demonstrate that the correct pose is estimated when multiple objects appear. It could also be extended to the case when the object is not visible by rejecting estimates with an high energy. In general, our method can be easily modified for certain applications, e.g., by including prior as we did for humans. Other possibilities are multi-cue integration and exploitation of an hierarchical structure, however, these features are object specific and not suitable for a general solution.

References

1. Bray, M., Kohli, P., Torr, P.: Posecut: Simultaneous segmentation and 3d pose estimation of humans using dynamic graph-cuts. In: European Conf. on Computer Vision. Volume 3952 of LNCS., Springer (2006) 642–655
2. Brox, T., Rosenhahn, B., Cremers, D., Seidel, H.P.: High accuracy optical flow serves 3-d pose tracking: Exploiting contour and flow based constraints. In: European Conf. on Computer Vision. Volume 3952 of LNCS., Springer (2006) 98–111

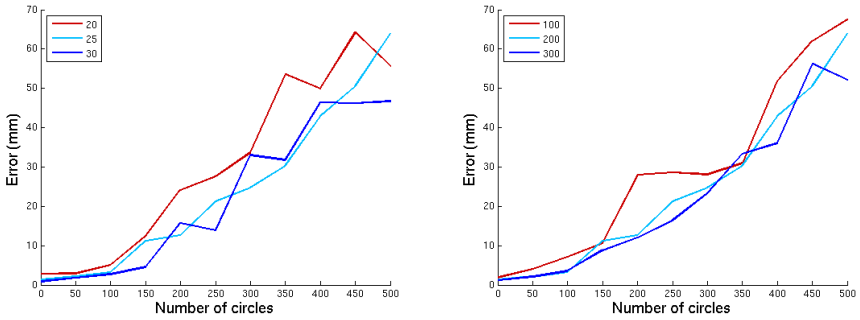

Fig. 5. Average error of the estimates for different numbers of iterations and 200 particles *(left)* and for different numbers of particles and 25 iterations *(right)*. 200 particles and 25 iterations are sufficient for rigid bodies.

Fig. 6. Estimates for 12 poses from a motion sequence from the CMU database (The estimated poses of the human model are projected onto the silhouette images). Each row shows one of the three views.

3. Rosenhahn, B., Brox, T., Weickert, J.: Three-dimensional shape knowledge for joint image segmentation and pose tracking. Int. J. of Computer Vision **73**(3) (2007) 243–262
4. Lowe, D.: Three-dimensional object recognition from single two-dimensional images. Artificial Intelligence **31**(3) (1987) 355–395
5. Lowe, D.: Fitting parameterized three-dimensional models to images. IEEE Trans. on Pattern Analysis and Machine Intelligence **13**(5) (1991) 441–450
6. Ansar, A., Daniilidis, K.: Linear pose estimation from points or lines. IEEE Trans. on Pattern Analysis and Machine Intelligence **25**(5) (2003) 578–589
7. Gall, J., Rosenhahn, B., Seidel, H.P.: Robust pose estimation with 3d textured models. In: IEEE Pacific-Rim Symposium on Image and Video Technology. Volume 4319 of LNCS., Springer (2006) 84–95
8. Lepetit, V., Pilet, J., Fua, P.: Point matching as a classification problem for fast and robust object pose estimation. In: IEEE Conf. on Computer Vision and Pattern Recognition. Volume 2. (2004) 244–250
9. Besl, P., McKay, N.: A method for registration of 3-d shapes. IEEE Trans. on Pattern Analysis and Machine Intelligence **14**(2) (1992) 239–256
10. Zhang, Z.: Iterative point matching for registration of free-form curves and surfaces. Int. J. of Computer Vision **13**(2) (1994) 119–152
11. Gall, J., Potthoff, J., Schnörr, C., Rosenhahn, B., Seidel, H.P.: Interacting and annealing particle systems – mathematics and recipes. J. of Mathematical Imaging and Vision (2007) To appear.
12. Kirkpatrick, S., Jr., C.G., Vecchi, M.: Optimization by simulated annealing. Science **220**(4598) (1983) 671–680
13. Doucet, A., de Freitas, N., Gordon, N., eds.: Sequential Monte Carlo Methods in Practice. Springer, New York (2001)
14. Gall, J., Rosenhahn, B., Seidel, H.P.: An Introduction to Interacting Simulated Annealing. In: Human Motion - Understanding, Modeling, Capture and Animation. Springer (2007) To appear.
15. Douc, R., Cappe, O., Moulines, E.: Comparison of resampling schemes for particle filtering. In: Int. Symposium on Image and Signal Processing and Analysis. (2005) 64–69
16. Moral, P.D.: Feynman-Kac Formulae. Genealogical and Interacting Particle Systems with Applications. Springer, New York (2004)
17. Stolfi, J.: Oriented Projective Geometry: A Framework for Geometric Computation. Academic Press, Boston (1991)
18. Murray, R., Li, Z., Sastry, S.: A Mathematical Introduction to Robotic Manipulation. CRC Press, Boca Raton, FL (1994)
19. Pennec, X., Ayache, N.: Uniform distribution, distance and expectation problems for geometric features processing. J. of Mathematical Imaging and Vision **9**(1) (1998) 49–67
20. Felzenszwalb, P., Huttenlocher, D.: Distance transforms of sampled functions. Technical Report TR2004-1963, Cornell Computing and Information Science (2004)
21. Gall, J., Rosenhahn, B., Brox, T., Seidel, H.P.: Learning for multi-view 3d tracking in the context of particle filters. In: Int. Symposium on Visual Computing (ISVC). Volume 4292 of LNCS., Springer (2006) 59–69
22. Brox, T., Rosenhahn, B., Kersting, U., Cremers, D.: Nonparametric density estimation for human pose tracking. In: Pattern Recognition (DAGM). Volume 4174 of LNCS., Springer (2006) 546–555
23. CMU: Graphics lab motion capture database http://mocap.cs.cmu.edu/.